Une Analyse préalable à l'indexation de transcriptions de conversations téléphoniques

Caroline Tambellini, Catherine Berrut, Christophe Brouard

Une Analyse préalable à l'indexation de transcriptions de conversations téléphoniques

Caroline Tambellini, Catherine Berrut, Christophe Brouard

Laboratoire CLIPS-IMAG
BP 53
38041 Grenoble cedex 9
caroline.tambellini@imag.fr, catherine.berrut@imag.fr,
christophe.brouard@imag.fr

RÉSUMÉ. *Nous nous intéressons dans cet article au problème de l'indexation de documents audio de type « conversation téléphonique ». Nous nous interrogeons en particulier sur le bien fondé de l'utilisation, pour ce type de documents, des méthodes d'indexation classiquement utilisées en recherche d'information textuelle. Pour répondre à ces questions, nous revisitons certaines hypothèses de la recherche d'information en étudiant la spécificité et l'applicabilité de ces hypothèses à des transcriptions de conversations téléphoniques. Nos observations nous conduisent à proposer les bases d'un système d'indexation pour ce type de documents qui comprend un module de découpage thématique de la conversation.*

ABSTRACT. *This paper deals with the problem of indexing the "phone conversation audio documents. In particular, we want to know how to use, for this kind of documents, the indexing methods traditionally used in textual information retrieval. To answer these questions, some information retrieval assumptions are revisited. The applicability and the specificity of these assumptions to phone conversations transcriptions are studied. Our study makes it possible to propose the bases of an indexing system for kinds type of documents which includes a module of conversation topic segmentation.*

MOTS-CLES : *Recherche d'information, indexation de documents audio, découpage thématique, analyse de conversations*

KEYWORDS : *Information retrieval, audio documents indexing, topic segmentation, conversation analysis*

1. Introduction / Problématique

L'émergence des nouvelles technologies de communication (conférences téléphoniques, vidéo conférences, …) génère de nouveaux besoins en terme de recherche d'information. Ainsi, par exemple, dans le cadre des entreprises virtuelles (entreprises dont les membres sont répartis géographiquement à différents endroits) le couplage des moyens de communication actuels (permettant notamment la réunion téléphonique) à des outils dédiés qui apportent un support à la communication, tels que des systèmes de recherche d'information en ligne, apparaît comme une nécessité.

Les finalités de systèmes de recherche d'informations pour la conférence téléphonique peuvent se situer à trois niveaux . Durant la réunion téléphonique, on peut s'assurer que l'ordre du jour est bien suivi. Pour cela, une correspondance entre le contenu de l'ordre du jour et les propos échangés doit être établie.

Il peut aussi être profitable de mettre à disposition des différents interlocuteurs les documents en rapport avec les thèmes abordés. Ces documents proposés aux interlocuteurs peuvent être trouvés dans une base de documents générale ou peuvent être mis à disposition par un des intervenants. Ainsi si un interlocuteur souhaite approfondir ou vérifier un point, il peut le faire grâce aux documents qu'il aura à sa disposition.

Dans un souci d'efficacité, ces tâches doivent être effectuées au fil de la conversation. Il existe donc une contrainte de temps que l'on ne rencontre pas dans les systèmes de recherche d'information « classique ».

Après une réunion, il est intéressant de retrouver seulement des parties de la réunion traitant d'un thème donné. En effet, lors d'une réunion, plusieurs points sont souvent abordés et une personne extérieure (ou non d'ailleurs) à la réunion peut souhaiter connaître (ou revoir) ce qui a été dit sur un point précis. De même, la rédaction du compte rendu de la réunion est une tâche qui s'effectue après la réunion. La détermination de ces parties de la réunion ainsi que la rédaction du compte rendu de la réunion nécessitent un outil de découpage thématique de la réunion téléphonique.

Il s'agit donc de mettre en place un système de recherche d'information sur des données de type conversations téléphoniques, et avec des finalités spécifiques. Ce problème nous amène à nous poser les deux questions suivantes :

- Quelle est la spécificité des documents "conversation téléphonique", en regard des documents traditionnellement traités en recherche d'informations ?

- Quel positionnement avoir par rapport à une recherche d'information classique, c'est-à-dire basé sur des documents textuels consistants ?

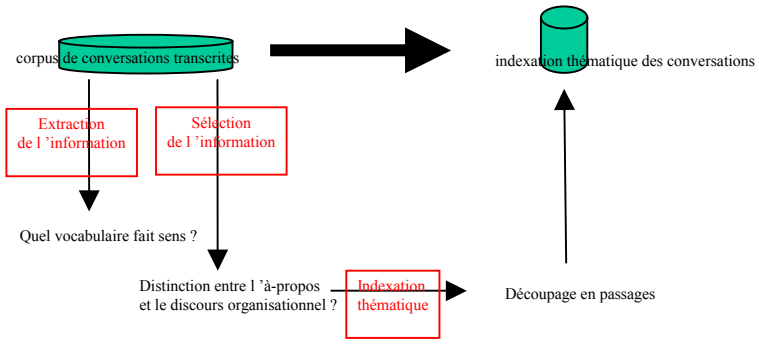

Figure 1. *Processus d'indexation thématique des conversations téléphoniques*

Pour ce faire, nous avons analysé un corpus de réunions téléphoniques transcrites. Cette analyse nous a amené à une réflexion sur l'indexation thématique des conversations (figure 1).

Tout système de recherche d'information "textuel" présuppose une "extraction" du vocabulaire qui fait sens. Et la question de l'applicabilité des hypothèses de RI qui sous-tendent ce processus d'extraction au cas des transcriptions de conversations, se pose. De plus, lorsqu'il s'agit de transcriptions de conversations, un processus de sélection visant à la distinguer l'"à propos" du "discours organisationnel" semble être un autre pré-requis incontournable à l'indexation thématique.

Le chapitre suivant présente un état de l'art dans le domaine de l'indexation de documents audio. Le chapitre 3 détaille le corpus de réunions téléphoniques que nous avons conçu et utilisé. Le chapitre 4 présente la méthode suivie dans l'analyse du corpus. Le chapitre 5 présente l'extraction et la sélection d'information que nous proposons en les mettant en perspective avec ce qui se pratique classiquement en recherche d'information. Enfin le chapitre 6 présente l'indexation thématique.

2. Les systèmes d'indexation de documents audio

Le traitement des documents « conversation téléphonique » ramène à deux problématiques que sont l'application de la recherche d'information à de l'audio (tâche d'indexation) et la segmentation thématique de documents. Nous nous intéressons à la combinaison de ces deux problématiques.

La conversation apporte des particularités qui peuvent être utilisées pour segmenter la conversation. Il n'est pas toujours évident de déterminer une unité de traitement pour le discours spontané. Dans certains cas, la dernière partie d'une phrase et la première partie de la phrase suivante forme un seul même intervalle de parole. Une façon de classifier les segments de discours peut se faire en 3

catégories : pause, silence, expressions interjectives et phrase de discours (Takagi *et al.*, 1996). Les caractéristiques prosodiques (intonations, hésitations, …) peuvent également être de bons indices pour découper le dialogue.

TDT (Topic Detection and Tracking) aborde la problématique de la segmentation thématique. TDT a pour objectif de déterminer automatiquement les passages d'une conversation traitant d'un même et seul sujet. TDT part de corpus de documents textuels en anglais, chinois et arabe. On peut découper TDT en cinq tâches techniques (Wayne, 2000) :

- **Segmentation** : trouver des régions thématiques homogènes.
- **Tracking** : trouver des passages additionnels à propos d'un thème donné.
- **Detection** : détecter et filtrer ensemble des nouveaux thèmes.
- **First Story Detection** : détecter de nouveaux thèmes.
- **Linking** : détecter quels passages traitent du même thème.

Cette problématique de segmentation thématique peut avoir de nombreuses applications. Par exemple, dans l'optique de supprimer le traitement manuel par des documentalistes des documents audiovisuels, le projet THEOREME a été mis en place en 1999 (RNRT, 1999). Dans ce projet, une des tâches est notamment la détection automatique de thèmes à partir d'une transcription entachée d'erreurs. Une des utilisations possibles des transcriptions d'émissions radio ou télévisées consiste en la possibilité d'analyser et de structurer automatiquement le contenu de tels documents. Comme pour les documents écrits on peut envisager l'automatisation de différentes tâches habituellement effectuées par des opérateurs humains, telles que l'indexation des documents, l'identification de thèmes et la détection d'événements.

De même, en 2001, le projet AUDIOSURF a vu le jour (AUDIOSURF, 2001). A partir d'un corpus d'enregistrements d'émission de radio de Radio France, le projet consiste à mettre en place un système permettant de gérer l'information audio selon des procédés du même type que ceux déjà mis en place pour les informations textuelles (indexation, recherche documentaire, filtrage, routage, extraction d'information, catégorisation, etc). Un tel système ne peut se limiter à la simple mise en relation des deux outils : reconnaissance de la parole et indexation de documents textuels. En effet, il faut tenir compte de la particularité des documents retranscrits, à savoir erreurs de transcription, absence de ponctuation, enrichissement par des données non textuelles comme le changement de locuteurs, etc. Il convient donc de penser correctement le lien entre les deux systèmes afin d'obtenir les meilleurs résultats.

TREC (Text REtrieval Conference) a également étudié la piste audio. TREC se base sur un corpus de 100h d'enregistrements d'émissions radio et télévisées pour TREC7 et 500 heures d'enregistrements pour TREC8. La tâche TREC Spoken Document Retrieval (SDR) (Garofolo *et al.*, 2000) a été créée pour résoudre le problème d'indexation des documents audio. Cette tâche s'effectue en deux temps. Premièrement, un système de reconnaissance de la parole est appliqué à un flux

audio et génère une transcription du document audio. Dans un deuxième temps, la transcription est indexée et recherchée par un système de recherche d'information. Nous constatons que des systèmes d'indexation basés sur des documents textuels sont utilisés pour les transcriptions des documents audio. Nous pouvons nous demander dans quelle mesure nous pouvons appliquer un tel système à des transcriptions de documents audio. Ces transcriptions de documents audio n'ont-ils pas des particularités propres ?

La problématique de l'indexation de l'audio et notamment des conversations téléphoniques est une tâche à laquelle nous nous intéressons plus particulièrement. N. Boufaden et al. (Boufaden *et al.*, 2002) se sont intéressés à cette problématique d'indexation des conversations téléphoniques. Leur approche est basée sur un découpage thématique utilisé pour faciliter l'extraction d'information à partir de conversations téléphoniques transcrites. Ils effectuent des expérimentations avec un modèle de Markov caché utilisant des informations de différents niveaux linguistiques, des marques d'extra-grammaticalités et les entités nommées comme source additionnelle d'information.

Le découpage thématique repose sur l'utilisation d'informations linguistiques et extralinguistiques pour détecter les changements de thèmes. Les informations linguistiques sont essentiellement :

- des mots tels que *ok, right, well* appelés **marques lexicales**
- des adverbes temporaux, conjonctions appelés **marques syntaxiques**
- le rôle du locuteur dans le développement du thème appelé **marques discursives**

De nombreuses études sur la recherche d'information et le découpage thématique sur de l'audio ont été réalisées mais aucune étude n'a été réalisée afin de voir si les transcriptions de documents audio peuvent être considérés comme des documents textuels « normaux », c'est-à-dire habituellement utilisés en recherche d'information. En effet, ces transcriptions de documents audio respectent-elles la loi de Zipf, suivent-elles la conjecture de Luhn, qui sont les bases sur lesquelles se positionnent les systèmes de recherche d'information ? Ce sont ces questions auxquelles nous essayons d'apporter une réponse dans cet article.

3. Le corpus audio

3.1. Construction d'un corpus

Nous avons construit notre propre corpus de conversations grâce à une collaboration avec l'équipe GEOD (Groupe d'Etude sur l'Oral et le Dialogue) (GEOD) du CLIPS et la société CALISTEL (CALISTEL). La construction de notre corpus s'est faite en deux temps. Dans un premier temps, nous avons recueilli les conversations audio qui ont eu lieu lors de réunions téléphoniques. Dans un second

temps, ces conversations ont été retranscrites par un professionnel des sciences du langage afin de constituer notre corpus de conversations retranscrites. De ce fait, les transcriptions comprennent beaucoup d'annotations (ton, hésitations, confusions, répétitions, etc.…) (annexe 1). Même si ce sont des transcriptions, on ne travaille pas sur des documents textuels, mais sur des conversations téléphoniques complètement transcrites (locuteur, tour de parole, hésitations, etc.…). Ainsi, notre étude ne porte pas sur le signal de la conversation mais bien sur une transcription « ASCII » de la conversation. On élude volontairement la problématique liée au traitement du signal dans une première approche.

3.2. Description du corpus audio

Le corpus dont nous disposons comporte 13 transcriptions manuelles de réunions téléphoniques. Chaque réunion est d'une durée moyenne de 45 minutes. Ces conversations sont regroupées en trois grandes catégories : réunion de projet, brainstorming et entretien d'embauche. Nous avons choisi de considérer trois types de réunions différents afin d'étudier si les spécificités de chaque réunion apportent des indices supplémentaires pour nos tâches de recherche d'informations. Nous avons établi le nombre de mots, le nombre de tours de parole, le nombre de mots par tour de parole pour chaque transcription de réunions. Le détail de ces statistiques se trouve en annexe 2.

Chacun des trois types de réunions a ses particularités. Ainsi, une réunion de projet a lieu en général en début et tout au long de la mise en œuvre d'un projet. Cette réunion regroupe en général un responsable de la réunion et un cercle d'intervenants qui sont tour à tour actifs ou passifs. En général, ce type de réunion débute par un tour de table où chaque participant fait le point sur l'état d'avancement personnel au sein du projet et expose parfois ses soucis. Ensuite, quelques points sont débattus et enfin l'équipe fait le point sur les futures tâches à accomplir.

La réunion de type brainstorming, quant à elle, a en général pour but la récolte de nombreuses idées en un minimum de temps. On pourrait assimiler la réunion de type brainstorming à un outil de « résolution de problème ». Dans ce type de réunion, tous les intervenants participent en même temps, chacun étant sur un pied d'égalité vis-à-vis des autres participants.

Enfin, la réunion de type entretien fait penser en premier lieu à l'entretien d'embauche mais ce peut être tout type d'entretien. Dans ce type de réunion, on trouve un jury (composé au minimum d'une personne) face à la personne qui passe l'entretien. En général, au sein du jury se trouve un président de jury qui peut être assimilé au leader du groupe. Souvent, celui-ci dirige l'entretien. Ensuite, plusieurs hypothèses sont possibles. Soit seul le président du jury pose les questions et les autres membres du jury (dans le cas d'un jury composé de n personnes) écoutent et

se contentent de juger sans poser de questions, soit chaque membre du jury participe à tour de rôle en posant des questions au candidat.

4. Méthode d'analyse du corpus

4.1. Spécificités des documents « conversations téléphoniques »

Les documents « conversations téléphoniques » ont plusieurs particularités :

4.1.1. Leur contenu relève de deux tâches très différentes :

- le discours organisationnel, qui est l'énoncé oral de la gestion de la réunion. Ce discours permet de faire les liens entre les différents sujets abordés, correspond aux formules de politesse, aux passages de parole d'une personne à l'autre, etc.

Exemple : « *Nous allons passer au point suivant »*

- l'à propos, qui est l'énoncé oral de l'objet effectif de la réunion. Ce discours correspond au développement des sujets tels que annoncés dans l'ordre du jour.

Exemple : *« Que pensez-vous de ce point ? »*

Cet aspect «double discours» est une spécificité des documents conversations.

4.1.2. Ils sont multi locuteurs, et chaque interlocuteur a un rôle identifié.

Un document de type « réunion téléphonique » est donc un document multi locuteurs. Chaque locuteur prend la parole à plusieurs reprises et pour traiter de plusieurs points différents. Le document est donc initialement découpé en locuteurs. C'est un découpage physique de la conversation.

Ainsi comme le montre la figure 2, le locuteur 3 a pris la parole au 3ème et 5ème tours de parole.

4.1.3. Une description logique

Parallèlement, toute réunion a un objet généralement décrit dans un ordre du jour. La description de la réunion selon cet ordre du jour donne ainsi une décomposition logique du document. Cette décomposition doit permettre la mise en évidence :

- Des parties des documents où est traité chaque point de l'ordre du jour.

Ainsi, comme le montre la figure 2, le document est découpé selon les trois points de l'ordre du jour et par exemple, le point 3 est abordé par le locuteur 3 puis par le locuteur 2.

- Des portions (phrases, mots) du document pertinentes pour chaque point de l'ordre du jour, telles que le montrent les *étoiles* de la figure 2.

Connaître cette décomposition logique du document « réunion » permet la mise en place d'applications telles que le suivi de l'ordre du jour, la rédaction de comptes-rendus,…

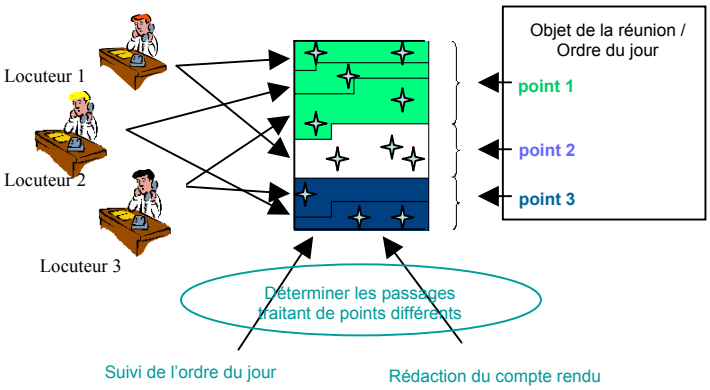

Figure 2. *Construction d'un document "conversation téléphonique"*

4.2. Démarche

Nous souhaitons construire une description logique associée à une réunion. Pour cela, nous devons mettre en place un découpage du document et nous verrons que cette étape est basée sur l'analyse du discours organisationnel. Parallèlement à cela, nous devons mettre en place une tâche d'extraction et de sélection du vocabulaire faisant sens dans le document.

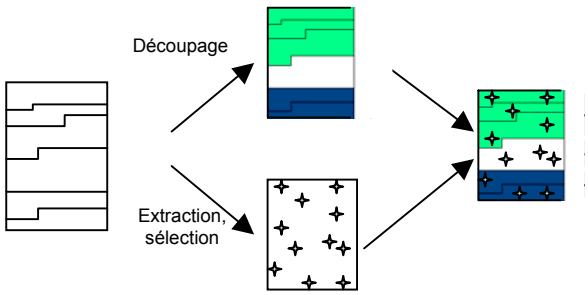

Figure 3. *Reconstitution du document logique*

Il nous est donc nécessaire de bâtir un processus permettant l'extraction, la sélection d'information ainsi que le découpage du document.

Lorsqu'il s'agit de documents textuels, ces processus sont connus et maîtrisés. Se pose alors la question de l'applicabilité de ces principes de la recherche d'information « textuelle » à des documents « conversation ». Et si oui, dans quelles mesures ?

Pour ce faire, nous proposons de suivre une démarche en trois étapes :

- ***vérification de l'applicabilité*** *:* il s'agit de vérifier si les hypothèses de la recherche d'information « textuelle » sont respectées avec les documents « conversation ».

- ***évaluation qualitative*** *:* dans quelle mesure ces hypothèses sont-elles applicables à nos documents « conversation » ?

- ***détermination de seuils*** *:* de quelle manière va-t-on appliquer ces hypothèses, quels sont les seuils, les métriques compatibles avec nos documents « conversation » ?

Notre méthode consiste donc, tout d'abord, à vérifier l'applicabilité des méthodes de Recherche d'information classique à chaque tâche. Si l'applicabilité s'avère possible, nous entreprenons ensuite l'évaluation qualitative. Dans un troisième temps, et si les deux premières étapes ont été concluantes, nous déterminons les seuils à appliquer à notre corpus « conversation ». A partir de là, nous appliquons cette démarche en trois étapes à chacun des trois processus : ainsi le chapitre 5 nous décrit cette démarche pour l'extraction et la sélection d'information, le chapitre 6 nous décrit, quant à lui, la façon dont nous proposons de découper logiquement du document.

5. Extraction et sélection d'information

5.1. Positionnement par rapport à la recherche d'information textuelle

Que ce soit pour le suivi de l'ordre du jour ou pour la rédaction du compte rendu d'une réunion, notre objectif peut se résumer dans la définition d'un processus d'indexation permettant une description thématique des conférences téléphoniques.

Nous pouvons présenter ce travail, comme un processus d'indexation avec deux grandes étapes : extraction et sélection, processus qui permettent à partir d'un corpus d'établir une représentation abstraite de son contenu.

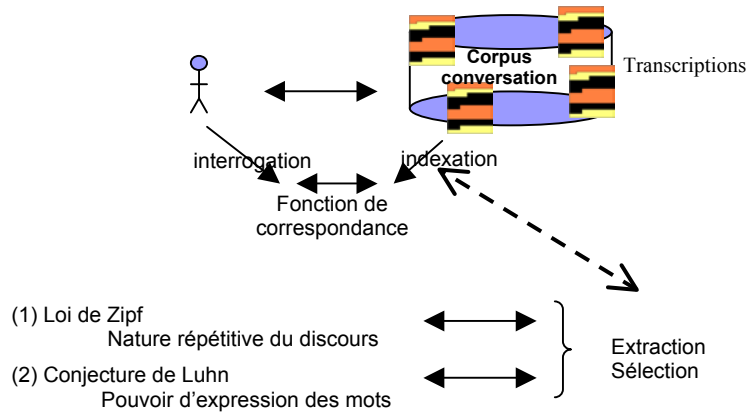

Figure 4. *Processus d'indexation*

Classiquement en recherche d'information, la tâche d'indexation d'un document textuel est basée sur la loi de Zipf et sur la conjecture de Luhn. L'utilisation de la loi de Zipf et de la conjecture de Luhn permettent la mise en place d'un processus d'indexation généralement en trois phases : extraction, sélection et pondération des termes d'indexation.

Comme le montre la figure 4, nous souhaitons vérifier que la loi de Zipf (partie 5.2.) et la conjecture de Luhn (partie 5.3.) sont applicables à des documents « conversation ».

5.2. Loi liée à l'occurrence des termes : loi de Zipf

5.2.1. La loi de Zipf

Le vocabulaire des documents classiquement utilisés en recherche d'information suit la loi de Zipf. La loi de Zipf dit que si l'on dresse une table de l'ensemble des mots différents d'un texte quelconque, classés par ordre de fréquence décroissante, la fréquence d'un mot est inversement proportionnelle à son rang dans la liste. Autrement dit, le produit de la fréquence de n'importe quel mot par son rang est constant, ce que traduit la formule :

rang du terme * (fréquence du terme / nombre de termes) = constante

Ainsi, l'analyse statistique des documents textuels en anglais montre que les mots les 20% les plus fréquents représentent 70% du vocabulaire des documents écrits (Salton, 1975).

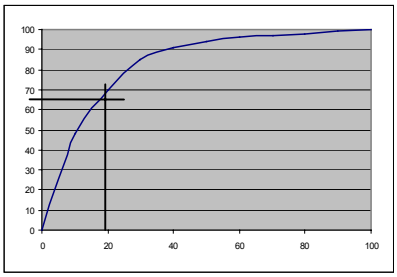

Figure 5. *Usage des mots dans les documents de RI*

5.2.2. Vérification de l'applicabilité de la loi de Zipf sur des documents « conversation »

Nous avons établi la fréquence des termes les plus courants dans chaque conversation. Ensuite, nous avons effectué le coefficient R * F/N, où R est le rang du terme, F est la fréquence du terme et N est le nombre total des termes.

Voici les résultats obtenus pour notre corpus de conversations :

	Brainstorming				Entretien				Réunion projet		
Rang	Terme	F	R * (F/N)	Rang	Terme	F	R * (F/N)	Rang	Terme	F	R * (F/N)
111	Projets	24	0,120	151	professionnel	4	0,143	137	coup	22	0,135
112	Après	24	0,121	152	prochaine	4	0,144	138	chose	22	0,136
113	Veux	23	0,117	153	poser	4	0,145	139	vais	21	0,130
114	Sont	23	0,118	154	parfait	4	0,146	140	clair	21	0,131
115	Idée	23	0,119	155	oh	4	0,147	141	avoir	21	0,132
,,,											
849	membres	2	0,077	246	transcription	2	0,117	921	partout	2	0,082
850	mélanger	2	0,077	247	toute	2	0,117	922	parties	2	0,082
851	meilleure	2	0,077	248	toujours	2	0,118	923	particulier	2	0,082
852	maximum	2	0,077	249	tombe	2	0,118	924	pars	2	0,083
853	marrant	2	0,077	250	terme	2	0,119	925	parle	2	0,083
,,,											
1202	Action	1	0,054	407	ambiance	1	0,097	1 265	accepter	1	0,056
1203	accueillent	1	0,054	408	aller	1	0,097	1 266	accélérer	1	0,057
1204	accessible	1	0,054	409	aider	1	0,097	1 267	accéder	1	0,057
1205	abstraction	1	0,054	410	agenda	1	0,097	1 268	absente	1	0,057
1206	Abord	1	0,054	411	activités	1	0,098	1 269	aborder	1	0,057

Figure *6.* *Illustration de la loi rang-fréquence pour le corpus des conversations*

On peut voir sur la figure 6 que l'on retrouve une constante pour le coefficient R*F/N. Cette constante est de l'ordre de 0,1 (si l'on fait une moyenne des valeurs).

5.2.3. Analyse qualitative de la loi de Zipf pour les documents « conversation »

Comme le montre la figure 7, en superposant la courbe sur l'usage des mots dans les documents « conversations » et celle de l'usage des mots dans les documents « classiques » de recherche d'information, nous constatons que l'on retrouve la

même tendance. Cette figure fait ressortir le caractère encore plus répétitif du vocabulaire de la conversation.

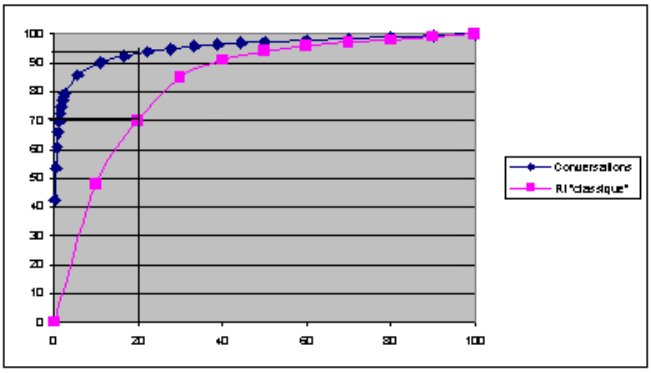

Figure 7. *Statistiques sur l'usage des mots dans l'ensemble des conversations de notre corpus*

Lorsque nous étudions les statistiques effectuées sur notre corpus de conversations, nous constatons que les mots les 10% les plus fréquents couvrent entre 70% et 80% du vocabulaire des conversations et nous constatons même que si nous faisons cette statistique sur l'ensemble des conversations, nous obtenons un taux de couverture du vocabulaire de près de 90%. Ceci nous montre que le vocabulaire des conversations est extrêmement répétitif.

On constate (voir figure 8) également que ce sont les transcriptions des réunions de type entretien d'embauche qui se rapprochent le plus de l'écrit.

Type de conversations	Taux de couverture du vocabulaire par les x% de mots les plus fréquents	
	10%	20%
Ecrit	≅ 50%	≅ 70%
Entretien d'embauche	≅ 70%	≅ 80%
Réunion de projet	≅ 80%	≅ 85%
Brainstorming	≅ 80%	≅ 90%
Toutes les conversations	≅ 90%	≅ 95%

+ proche de l'écrit

- proche de l'écrit

Figure 8. *Taux de couverture du vocabulaire pour les différents types de conversation*

5.2.4. Conclusion

Nous constatons que le vocabulaire des documents « conversation » respecte la loi de Zipf et nous permet d'établir une courbe sur la statistique d'usage des mots similaire à celle connue pour les documents textuels. Nous venons donc d'établir l'applicabilité et l'évaluation qualitative de la loi de Zipf sur les documents « conversation ».

5.3. Critère de sélection des termes pour l'indexation : conjecture de Luhn

5.3.1. La conjecture de Luhn

En recherche d'information, la conjecture de Luhn nous indique le pouvoir d'expression des mots dans un texte en fonction de leur fréquence. Ceci est représenté par la courbe de la figure 9.

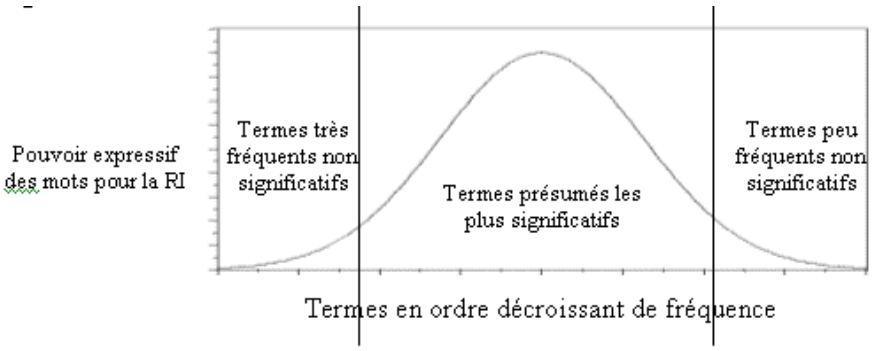

Figure 9. *Conjecture de Luhn : pouvoir d'expression des mots*

Sur cette courbe, en abscisse, les mots sont ordonnés du plus fréquent au moins fréquent. En ordonnée, cette courbe montre que les extrêmes (mots peu ou trop utilisés) offrent un pouvoir expressif limité, contrairement aux mots d'utilisation moyenne. Le pouvoir expressif des mots est ici basé sur le calcul de leur fréquence. Toutefois, comme nous allons le voir, d'autres méthodes peuvent être utilisées.

5.3.2. Autres citères de sélection

Parmi les systèmes indexant des documents textuels en français, un certain nombre (IOTA (Bruandet *et al.*, 1997), …) utilisent des indexations non pas basées sur des statistiques mais sur une analyse de la langue naturelle. En analyse morpho-syntaxique du langage, on distingue deux grandes familles de mots :

- une famille regroupant les adjectifs qualificatifs, les substantifs propres, les substantifs communs et les verbes (groupe 1 sur la figure 10)
- une autre regroupant le reste du vocabulaire (groupe 2 sur la figure 10).

Le vocabulaire porteur de sens pour la recherche d'information se situe dans le groupe 1, comme le montre la figure 10 (Palmer, 1990).

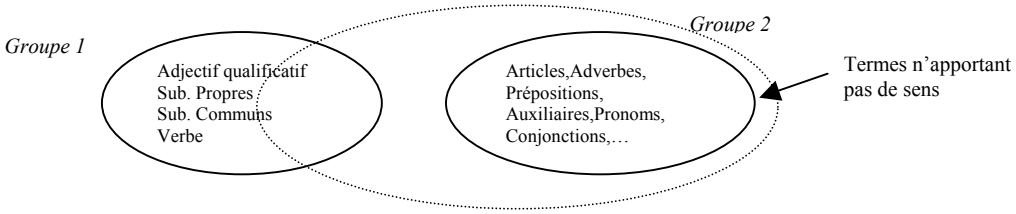

Figure 10. *Catégories de vocabulaire*

Nous avons donc étudié de façon qualitative et de façon similaire le vocabulaire utilisé dans les documents « conversation ».

5.3.3. Applicabilité de la conjecture de Luhn et détermination de seuils

Méthode

L'analyse de notre corpus fait ressortir le vocabulaire apportant le plus de sens est essentiellement constitué des substantifs communs. Les substantifs communs comprennent le vocabulaire relatif aux différents thèmes abordés. Il est donc intéressant de considérer cette famille de mots. Nous avons étudié la fréquence des catégories morpho-syntaxiques en fonction de la fréquence des mots (figure 11). Nous pourrions nous satisfaire d'une méthode de sélection basée sur les catégories morpho-syntaxiques mais pour des raisons d'efficacité, nous souhaitons vérifier l'applicabilité de la conjecture de Luhn qui permettrait une sélection plus rapide puisque ne requérant pas une analyse morpho-syntaxique mais simplement un seuil sur un nombre d'occurrences.

Résultats

Nous observons que les substantifs (qui d'après notre étude sont porteurs de sens) se situent entre deux seuils que nous avons déterminés (figures 11 et 12). On peut donc en déduire l'applicabilité de la conjecture de Luhn.

En abscisse, nous trouvons les termes par ordre décroissant de fréquence dans l'ensemble des conversations et en ordonnée, nous avons le nombre de termes de chaque catégorie.

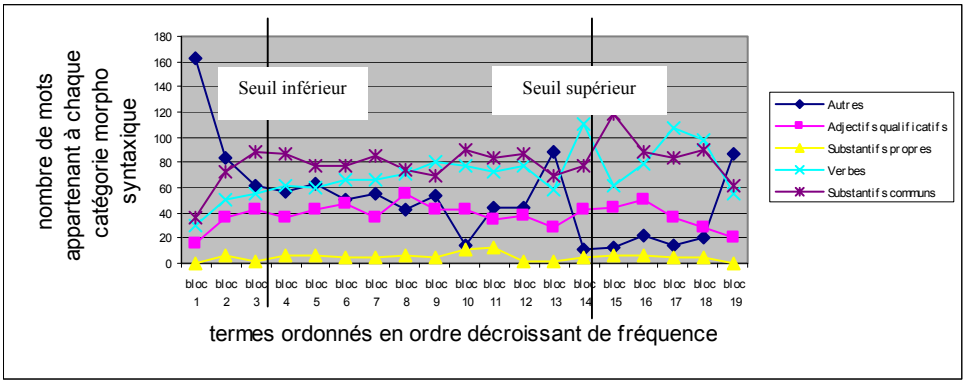

Figure 11. *Fréquence des différentes catégories de mots par portion de 250 termes*

En tenant compte des catégories intéressantes (les substantifs communs), les seuils optimaux sont les suivants :

- seuil inférieur : 3ème bloc (1 bloc = 250 mots)
- seuil supérieur : 14ème bloc

Entre ces deux valeurs seuils, nous voyons que la proportion de mots les plus fréquents appartiennent à la catégorie morpho-syntaxique des substantifs communs. Seuls les termes situés entre ces deux seuils seront donc pris en compte.

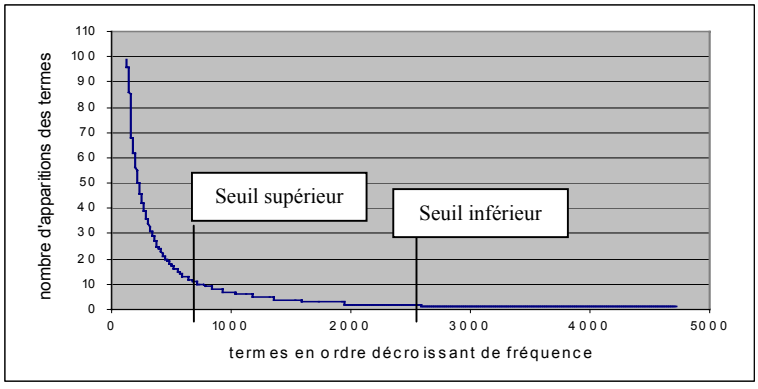

Figure 12. *Zoom sur le nombre d'apparitions des termes par ordre décroissant de fréquence*

De façon plus précise, la figure 12 nous permet de déterminer un seuil inférieur à 2 apparitions et un seuil supérieur à 10 apparitions d'un même mot.

Pour conclure sur ce chapitre, nous avons constaté l'applicabilité de la loi de Zipf et de la conjecture de Luhn, nous avons ensuite déterminé les seuils à appliquer

à notre corpus de documents « conversation » pour mettre en place l'extraction et la sélection d'information. La tâche de découpage thématique peut alors être étudiée.

6. Découpage thématique

6.1. Principe

Le but du découpage thématique consiste à identifier les parties logiques du document « conversation ». Pour ce faire, notre approche détermine les frontières (ou ruptures) entre les thèmes (relatifs à différents points de l'ordre du jour). Nous verrons que ces frontières s'expriment via le discours organisationnel des conversations.

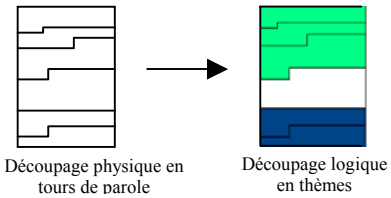

Figure 13. *Principe de découpage du document*

6.2. Indices de changements de thèmes

Différentes méthodes peuvent être utilisées pour faire une segmentation thématique d'un document (Callan, 1994), (Kaszkiel, 1997), (Moffat *et al.*, 1994), (Salton *et al.*, 1993). Nous nous positionnons dans une optique de suivi thématique positif et non négatif tel que le fait un système de TextTiling (Hearst *et al.*, 1993). En effet, le TextTiling cherche les ruptures de thèmes et les identifie lorsqu'un passage du document présente un moins grand nombre de mots traitant du thème. Dans notre approche nous privilégions le suivi thématique positif, c'est-à-dire que nous considérons que nous avons un changement de thème lorsque nous rencontrons une forte densité de mots que nous considérons être des mots de rupture. Ceci est possible grâce à la présence d'un discours organisationnel propre aux conversations téléphoniques.

6.3. Le découpage

Nous avons comparé un découpage effectué manuellement, avec un découpage calculé à partir du vocabulaire de conversation dédié explicitement au changement de thème (mots de rupture dans le dialogue, interrogations particulières, etc).

6.3.1. Vocabulaire de rupture, d'interrogations

Par une analyse du vocabulaire des conversations téléphoniques, nous avons mis en évidence des mots identifiant des zones de changement de thèmes. Ces mots sont réunis en trois catégories :

- *mots « d'approbation »* : ouais, d'accord, okay
- *mots de « changement »* : passons, passer, suite, suivant
- *question* : ?

Pour les expérimentations nous comparons les ruptures détectées manuellement aux ruptures détectées par le système par accumulation de ce vocabulaire.

La comparaison s'établit entre

- les résultats manuels
- la détection de tous les mots indiqués ci-dessus (appelé ensuite vocabulaire de rupture 1)
- la détection uniquement des mots de changement et des questions d'autre part (appelé ensuite vocabulaire de rupture 2).

6.3.2. Expérimentations

Ces expérimentations font ressortir qu'il existe un seuil dans le nombre de mots de rupture, au-dessus duquel les ruptures donnent une bonne approximation des résultats trouvés manuellement.

A partir de là, nous établissons une évaluation de ce seuil, et donc nous avons choisi de le trouver parmi les valeurs donnant les meilleurs résultats dans nos calculs. Pour ce faire, nous procédons en deux étapes :

- quel vocabulaire établit les meilleures détections de rupture (vocabulaire de rupture 1 ou 2) ?
- connaissant le meilleur vocabulaire, quelle valeur de seuil donne les meilleurs résultats ?

6.3.3. Résultats

Le tableau en annexe 3 donne pour chaque réunion, pour différents seuils, le taux de rappel et de précision lorsque nous utilisons le vocabulaire de rupture 1. La moyenne globale du rappel est de 0,37, alors que la précision moyenne est de 0,35. Le détail des résultats se trouve en annexe 3.

Comme le montre la figure 14, la méthode prenant en compte le vocabulaire de rupture 2 (les mots de changement + les points d'interrogation) donne un rappel moyen de 0,51 et une précision moyenne de 0,59.

	Seuil	pertinents	retrouvés	pert retrouvés	rappel[1]	précision[2]
Brainstorming 1	1	6	9	5	0,83	0,56
	2	6	3	3	0,50	1,00
	3	6	2	2	0,33	1,00
Brainstorming 2	1	4	7	2	0,50	0,29
	2	4	4	2	0,50	0,50
	3	4	3	2	0,50	0,67
Brainstorming 3	1	6	12	3	0,50	0,25
	2	6	4	2	0,33	0,50
	3	6	2	2	0,33	1,00
Brainstorming 4	1	6	13	4	0,67	0,31
	2	6	5	3	0,50	0,60
	3	6	2	1	0,17	0,50
Brainstorming 5	1	4	7	4	1,00	0,57
	2	4	4	2	0,50	0,50
	3	4	3	2	0,50	0,67
moyenne					0,51	0,59
moyenne des meilleurs rappels - précisions					*0,70*	*0,79*
moyenne avec le seuil de 1					*0,70*	*0,39*
moyenne avec le seuil de 2					*0,47*	*0,62*
moyenne avec le seuil de 3					*0,37*	*0,89*

Figure 14. *Rappel et précision pour différents seuils en utilisant le vocabulaire de rupture 2 (les mots de changement + les points d'interrogation)*

La méthode donnant les meilleurs rappels et précisions moyens est donc la méthode prenant en compte le vocabulaire de rupture 2. Notons que l'on peut avoir

[1] Rappel = nombre de pics pertinents trouvés / nombre de pics pertinents
[2] Précision = nombre de pics pertinents trouvés / nombre de pics trouvés

des précisions de 1 mais dans ce cas là, le rappel est très faible, ceci est dû au fait que l'on prend très peu de pics en compte.

A partir des résultats obtenus, nous avons déterminé le seuil optimal pour les cinq brainstormings avec ce vocabulaire (voir les dernières lignes de la figure 14). Un seuil de 2 est une bonne optimisation.

Le tableau (figure 15) donne pour chaque réunion, pour différents seuils, le taux de rappel et de précision lorsque nous utilisons le vocabulaire de rupture 2. La moyenne globale du rappel est de 0,51, alors que la précision moyenne est de 0,59.

	Seuil	pertinents	etrouvés	Pert retrouvés	Rappel	Précision
Brainstorming 1	2	6	3	3	0,50	1,00
Brainstorming 2	2	4	4	2	0,50	0,50
Brainstorming 3	2	6	4	2	0,33	0,50
Brainstorming 4	2	6	5	3	0,50	0,60
Brainstorming 5	2	4	4	2	0,50	0,50
Moyenne					0,47	0,62

Figure 15. *Bilan en utilisant le vocabulaire de rupture 2 (les mots de changement + les points d'interrogation) et un seuil égal à 2, pour les brainstormings*

6.4. Interprétation

Si nous regardons les ruptures données par la combinaison des mots de « changement » et des « questions », nous constatons que les ruptures détectées manuellement sont également détectées par le système par une plus grande densité de ces mots de « changement » et « questions». Au contraire, les mots « d'approbation » ne semblent pas être porteurs de changements de thèmes. La combinaison mots de « changement » + « question » est donc la meilleure pour notre cas.

Bien évidemment, on constate que d'autres ruptures sont détectées alors qu'elles n'avaient pas été faites par un utilisateur. Ceci peut s'expliquer en partie par le fait qu'il n'est pas toujours évident de déterminer des ruptures de thèmes manuellement dans une conversation car on ne passe pas de façon nette et précise d'un thème à un autre, comme on pourrait le faire dans un texte écrit.

Toutefois, nous retrouvons, dans chacune des conversations, un pic bien marqué lors du changement de thème entre le thème du poster et celui des vacances.

Ces premières expérimentations suggèrent que les indices de rupture présentent un réel intérêt pour le découpage de la conversation.

Il paraîtrait intéressant de combiner maintenant ces indices avec des indices tel que le rôle de la personne prononçant ces mots de « rupture ».

Ces expérimentations confirment également l'idée selon laquelle il est difficile et peu approprié de découper le document de façon stricte. En effet, on ne change pas de thème dans une conversation de façon nette et précise. Et souvent, deux thèmes se « chevauchent » dans la conversation.

7. Conclusion

Notre étude répond aux deux questions suivantes :

- Quelle est la spécificité des documents "conversation téléphonique", en regard des documents traditionnellement traités en recherche d'informations ?

- Quel positionnement avoir par rapport à une recherche d'information classique, c'est-à-dire basé sur des documents textuels consistants ?

De ce fait, nous avons étudié l'adaptabilité et l'applicabilité des systèmes de recherche d'information textuels aux documents « conversation téléphonique ». Pour ce faire, nous avons fait une étude qualitative du vocabulaire utilisé dans les conversations téléphoniques en regard de celui utilisé dans les documents textuels « classiques ». Cette étude nous permet d'avoir un premier système permettant l'indexation thématique de transcriptions de conversation.

Le système se décompose entre extraction, sélection et découpage thématique comme le montre la figure 16.

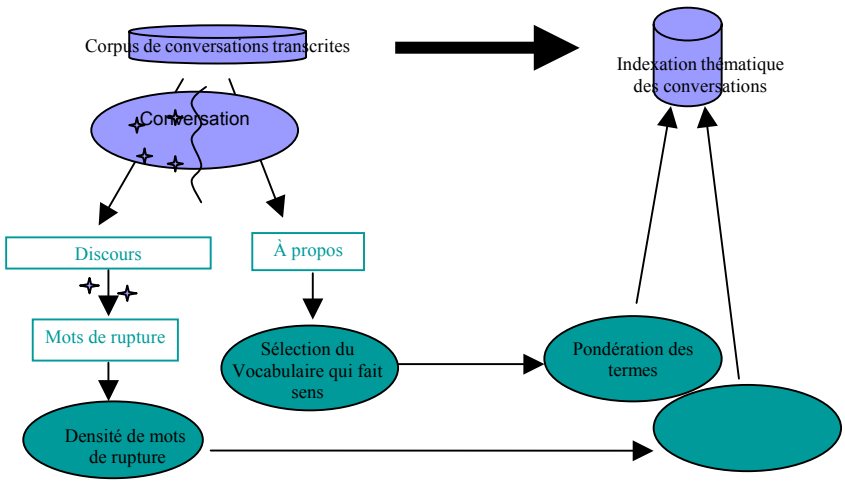

Figure 16. *Schéma récapitulatif de la démarche d'indexation thématique des conversations*

Comme le montre la figure 16, un document de type 'conversation' est composé d'un double discours : le discours organisationnel et l'à propos. L'étude du discours organisationnel permet d'en extraire des mots de rupture qui servent au découpage en fenêtre thématique du document. Le vocabulaire de l'à propos, quant à lui, est utilisé pour sélectionner le vocabulaire faisant sens. La combinaison de l'application d'une pondération simple (par exemple tf*idf) à ces termes (de l'à propos) faisant sens et du découpage en fenêtre du document permet une indexation thématique des conversations.

La prochaine tâche à accomplir est d'intégrer l'aspect signal de la conversation que nous avons volontairement éludé dans un premier temps, dans un souci de simplification. Nous souhaitons mettre en place cette étape à court terme. A plus long terme, nous bâtirons les applications de recherche, filtrage d'informations, aide à la rédaction de compte-rendu, etc.

Nous avons essentiellement basé notre étude sur le vocabulaire des conversations mais la conversation contient bien d'autres indices. Comme nous l'avons vu le rôle du locuteur va tenir une place importante car on peut notamment penser que c'est la personne menant l'entretien qui va amener les changements de thèmes. D'autres indices contextuels propres au caractère « situé » des conversations téléphoniques apportent également de l'information. Tous ces éléments doivent coopérer pour bâtir un système robuste dédié à la réunion téléphonique.

8. Annexes

Annexe 1 : Extrait de transcription de conversation téléphonique

; CDR: 00.00

; TRV: 00.00

; File: Bg2

; Last changes made on 12/02/2002

; Transcriber: ACD

; Comments:

;

xxxxBG2_1_0077_ACD_00: super <Laugh> . <B> bon eh bien [bé] <uh> <hes> peut-être que le mieux ce serait de commencer +/ pa= /+ <uh> d'abord +/ par par /+ le haut +/ du /+ +/ du /+ du poster donc c'est-à-dire +/ le /+ le slogan . est-ce que vous auriez <uh> des commentaires ou +/ des /+ des suggestions , peut-être des critiques <uh> à faire ?

xxxxBG2_1_0078_EM_00: mhm .

xxxxBG2_1_0079_BM_00: <uh> alors moi j'ai +/ peut-être une /+ <uh> peut-être une suggestion avant c'est , je voudrais savoir à qui est destiné le poster ? parce que suivant +/ la /+ la cible visée il n'aura +/ pas le même /+ <uh> pas le même discours .

xxxxBG2_1_0080_ACD_00: mhm .

xxxxBG2_1_0081_JCD_00: ouais .

xxxxBG2_1_0082_EM_00: mhm .

xxxxBG2_1_0083_ACD_00: oui .

xxxxBG2_1_0084_EM_00: <%> c'est vrai ouais .

xxxxBG2_1_0085_ACD_00: +/ ouais /+ ouais c'est vrai .

xxxxBG2_1_0086_BM_00: donc est-ce qu'il est grand public ou est-ce qu'il est <uh> plutôt à un public <uh> averti ? et peut-être que ça orientera <uh> le <*T>t

xxxxBG2_1_0087_JCD_00: mhm .

xxxxBG2_1_0088_ACD_00: oui bien sûr ouais . c'est vrai que là par exemple , celui-là -/ il est vraiment /- il a été fait pour un grand public <B> mais <uh> bon a priori ce serait plus pour faire +/ un /+ un vrai poster +/ bien /+ bien sérieux bon à présenter +/ aux /+ aux autres équipes , à d'autres labos <uh> internationaux ou nationaux donc <uh> <B> ce serait plus +/ un /+ un public averti .

xxxxBG2_1_0089_JCD_00: mhm .

xxxxBG2_1_0090_JCD_00: mhm .

xxxxBG2_1_0091_BM_00: d'acco= <*T>t

xxxxBG2_1_0092_EM_00: ah oui d'accord . parce que +/ dans la /+ +/ dans nos /+ dans nos consignes <uh> disons , on disait que ce poster c'était pour <uh> présenter le CLIPS aux futures conférences ou fêtes de la science ou colloques donc là +/ ça /+ c'est trop large là .

xxxxBG2_1_0093_JCD_00: +/ oui /+ oui c'est incompatible .

xxxxBG2_1_0094_ACD_00: <B> oui mais pourquoi <*T>t

xxxxBG2_1_0095_JCD_00: ouais .

xxxxBG2_1_0096_ACD_00: tu penses ? +/ pourquoi /+ pourquoi <*T>t

xxxxBG2_1_0097_EM_00: +/ non /+ non mais si tu dis <uh> ce qui est très juste <uh> que c'est plutôt orienté vers un public averti donc ça cadre bien avec conférence , colloque . par contre fête de la science bon c'est vrai que c'est peut-être un petit peu plus grand public quoi <%> <*T>t

xxxxBG2_1_0098_BM_00: voilà .

xxxxBG2_1_0099_ACD_00: mhm .

xxxxBG2_1_0100_BM_00: tout à fait <%> <*T>t

xxxxBG2_1_0101_ACD_00: <B> mais pourquoi -/ on ne serait pas /- on ne pourrait pas être plus clair <uh> . je veux dire <uh> pourquoi on ne serait pas capable de faire +/ un /+ <hm> <B> un poster +/ qui /+ qui puisse <uh> coller +/ aux /+ aux deux cas de figure .

Annexe 2 : Statistiques sur les mots pour chaque conversation

Dialogues	Nombre de mots	Nombre de tours de parole	Nombre de mots par tour de parole
Brainstorming1	8373	801	10,45
Brainstorming2	7892	991	7,96
Brainstorming3	8261	927	8,91
Brainstorming4	7226	993	7,28
Brainstorming5	6083	562	10,82
Entretien1	2793	207	13,49
Entretien2	1552	138	11,25
Entretien3	2304	171	13,47
Entretien4	2687	185	14,52
Réun. Projet 1	9142	723	12,64
Réun. Projet 2	7671	470	16,32
Réun. Projet 3	13228	1024	12,92
Réun. Projet 4	9079	718	12,64

Annexe3 : Rappel et précision pour différents seuils en utilisant le vocabulaire de rupture 1 (les mots d'approbation + les mots de changement + les points d'interrogation)

	Seuil	pertinents	retrouvés	pert retrouvés	Rappel	Précision
Brainstorming 1	5	6	11	4	0,67	0,36
	6	6	8	3	0,50	0,38

	Seuil	pertinents	retrouvés	pert retrouvés	Rappel	Précision
	7	6	6	2	0,33	0,33
Brainstorming 2	4	4	15	2	0,50	0,13
	5	4	6	2	0,50	0,33
	6	4	3	1	0,25	0,33
Brainstorming 3	6	6	9	3	0,50	0,33
	7	6	6	2	0,33	0,33
	8	6	4	2	0,33	0,50
Brainstorming 4	4	6	8	3	0,50	0,38
	5	6	3	1	0,17	0,33
	6	6	1	1	0,17	1,00
Brainstorming 5	4	4	14	1	0,25	0,07
	5	4	10	1	0,25	0,10
	6	4	3	1	0,25	0,33
Moyenne					0,37	0,35

9. Bibliographie

AUDIOSURF :
http://www.telecom.gouv.fr/rntl/AAP2001/Fiches_Resume/AUDIOSURF.htm

Boufaden, Lapalme, Bengio, *Découpage thématique des conversations : un outil d'aide à l'extraction*, TALN 2002, 24-27 juin 2002

Bruandet M-F., Chevallet J.P., Paradis F., *Construction de thesaurus dans le systeme de recherche d'information IOTA : application a l'extraction de la terminologie*, , in 1eres Journees Scientifiques et Techniques du Reseau Francophone de l'Ingerierie de la Langue de l'AUPELF-URF , Avignon - France, pp537-544, 15-16 Avril, 1997.

CALISTEL : http://www.calistel.com/

Callan ,*Passage-level evidence in document retrieval,* Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 1994

Garofolo J.S., Auzanne C.G., Voorhees E.M., *The TREC Spoken Document Retrieval Track : a success story*, 2000

GEOD : http://www-clips.imag.fr/geod/

Hearst Marti A., Plaunt Christian, *Subtopic structuring for full-length document access*, Proceedings of the Sixteenth Annual International ACM SIGIR conference on Research and Development in Information Retrieval, 1993, pp. 59-68

Kaszkiel Marcin, Zobel Justin, *Passage retrieval revisited*, SIGIR 1997

Moffat, Sacks-Davis, Wilkinson, Zobel, *Retrieval of partial documents*, Text REtrieval Conference 1994

Palmer Patrick, *Etude d'un analyseur de surface de la langue naturelle: application à l'indexation automatique de textes*, Ph.D. thesis, Université Joseph Fourier, 1990.

RNRT : http://www.telecom.gouv.fr/rnrt/projets/res_d97_ap99.htm

Salton, *A theory of indexing*, 1975

Salton, Allan, Buckley, *Approaches to passage retrieval in full text information systems*, ACM-SIGIR 1993

Takagi Kazuyuki, Itahashi Shuichi, *Segmentation of spoken dialogue by interjections, disfluent utterances and pauses*, 1996

Wayne Charles L., *Multilingual topic Detection and Tracking : Successful Research Enabled by Corpora and Evaluation*, 2000